
Articles

2023

Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI

na li

Technological University Dublin, d19125334@mytudublin.ie

Robert J. Ross

Technological University Dublin

Follow this and additional works at: <https://arrow.tudublin.ie/creaart>



Part of the [Cognitive Science Commons](#), and the [Engineering Commons](#)

Recommended Citation

Na Li and Robert Ross. 2023. Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23), March 13–16, 2023, Stockholm, Sweden. ACM, New York, NY, USA, 11 pages. DOI: 10.1145/3568162.3576999

This Conference Paper is brought to you for free and open access by ARROW@TU Dublin. It has been accepted for inclusion in Articles by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie.



This work is licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 4.0 License](#)
Funder: Science Foundation Ireland

Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI

Na Li
na.li@tudublin.ie

Technological University Dublin
Dublin, Ireland

Robert Ross
robert.ross@tudublin.ie

Technological University Dublin
Dublin, Ireland

ABSTRACT

Our research seeks to develop a long-lasting and high-quality engagement between the user and the social robot, which in turn requires a more sophisticated alignment of the user and the system than is currently commonly available. Close monitoring of interlocutors' states, and we argue their confusion state in particular, and adjusting dialogue policies based on this state of confusion is needed for successful joint activity. In this paper, we present an initial study of a human-robot conversation scenarios using a Pepper robot to investigate the confusion states of users. A Wizard-of-Oz (WoZ) HRI experiment is illustrated in detail with stimuli strategies to trigger confused states from interlocutors. For the collected data, we estimated emotions, head pose, and eye gaze, and these features were analysed against the silence duration time of the speech data and the post-study self-reported confusion states that are reported by participants. Our analysis found a significant relationship between confusion states and most of these features. We see these results as being particularly significant for multimodal situated dialogues for human-robot interaction and beyond.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; **Interaction design process and methods**.

KEYWORDS

confusion detection, user engagement, situated dialogues, woZ

ACM Reference Format:

Na Li and Robert Ross. 2023. Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*, March 13–16, 2023, Stockholm, Sweden. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3568162.3576999>

1 INTRODUCTION

Human-Robot Interaction (HRI) is a complex interdisciplinary field that straddles topics as varied as human-computer interaction, robotics, artificial intelligence, design, and philosophy [4]. It has been shown that systematic HRI study can improve effective communication in a variety of multitasking domains, e.g., digital learning environments [36], domestic environments [19], laboratory environments [32], and noisy and unpredictable environments [19].

Although HRI can leverage interaction across many different modalities, we argue that the conversational channel is of critical importance for naturalistic communication. To ensure users continually engage in interaction with robots, building a smooth and fluid conversation is, however, a challenging task, as it requires the system to provide appropriate responses to an interlocutor's nonverbal and verbal behaviours, mental states, and emotions. Indeed, emotions in people can modulate their behaviour in ongoing experience [4, chapter 8]. Therefore, designing emotional communication in HRI is critical, although there are many challenges in current emotion studies. For example, data collected for emotion recognition often comes from actors [7, 8], plus there is a lack of research related to engagement estimation not only in conversational HRI but also in spontaneous conversation [5, 6].

While emotion monitoring and estimation of engagement are important areas for research, we have identified confusion as one aspect of dialogue-centric HRI interaction that is particularly in need of systematic modelling. Compared to a level of "raw" emotion or engagement, confusion is a dynamic mental state that can be associated with both positive qualities such as interest, and negative qualities such as boredom and subsequent disengagement [12]. Being able to detect the state of confusion of an interlocutor in real time will likely improve engagement levels as we can adjust dialogue policy in response to confusion states in joint activities [10, 44]. While there have been limited works studying confusion states in general, with most works to date focused on the area of online learning, there has been very little research in the area for conversational systems and, in particular, little to nothing in the area of HRI. Given the above, in this paper, we present an exploratory study of confusion states in situated task-oriented HRI. More specifically, we address the following two research questions: (1) Do participants have a self-awareness that they are confused in a specific confusing situation in a controlled situated HRI? (2) What are the different manifest behaviours of the participants that we can detect when a user is in a confused state?

To address these issues, according to the definition of confusion that we defined [25], a WoZ [41] study was designed in HRI, with a subsequent feature analysis in the collected data. The contributions of this paper include the particular study design for confusion analysis, the insights derived from data analysis, and our analysis of the existing literature with respect to confusion detection. Finally, an anonymised feature set lifted form of this data¹ is made available for public analysis.



This work is licensed under a Creative Commons Attribution International 4.0 License.

¹<https://github.com/nalibjchn/HRI-FeatureData>

2 RELATED WORK

In this section, we briefly highlight some aspects of multimodal emotion detection, engagement estimation, and confusion detection that our work either builds on or has a relationship with.

2.1 Emotion Recognition

If a person has a strong ability to observe others' emotions and manage their own emotions, they are likely to contribute more successfully to an interaction with others [38]. Similarly, a social robot is arguably expected to have human-like capabilities of observing and subsequently predicting human emotion. Building on this idea, Spezialetti et al. [47] identifies three broad sets of tasks required to equip robots with emotional capabilities: (a) designing robot emotional states in existing cognitive architectures or emotional models; (b) formulating rich emotional expressions for robots through facial expression, gesture, voice, *etc.*; and (c) detecting and predicting human emotions. The first two broad sets are in robots oriented research that focusses on designing robots' behaviours to express robots' emotions in HRI, so that interlocutors can interpret the robots' emotions and even adopt their emotions [50]; while the last broad set is in human-centred research that focuses on designing an affect recognition system or model for a robot that is designed to observe user behaviour, which is also related to our study. Moreover, Cohn [9] indicated that human emotions cannot be observable because emotion is a cognition, a feeling, or a physiological and neuromuscular change. Therefore, he points out that emotion must be explained through an interaction context, a user survey, behaviours or physiological indicators [48].

Although emotion is not easily detectable directly, emotion and cognitive state can be indirectly observed. Facial expression is a natural emotional expression for a human being, as such many systems and models have been developed over the last few decades to label emotion from facial expressions. Facial Action Coding System (FACS) with AUs (facial action units) [9, 29] is one of those early models. There are also multiple examples of deep networks and in particular convolutional neural networks (CNNs) being applied directly for emotion recognition [40]. Recently, the "Dialogue Emotion Correction Network (DECN)" [26] has been proposed as a new correction model that includes an utterance-based emotion recognition engine alongside a conversation-based correction model.

Beyond facial expression, additional nonverbal signals that can suggest emotion are head pose and eye gaze. Liang et al. [27]'s experiments have verified the effects of gaze direction on perception of emotional expression. Their results suggest that with neutral or smiling faces, a direct gaze increases the possibility of users perceiving happy emotions, whereas avoidant gaze may increase the possibility of perceiving anger and fear. Moreover, head-pose estimation is inherently related to visual eye-gaze estimation. For example, Murphy-Chutorian and Trivedi [34] clarified that people with different head poses can reflect more emotional information (*i.e.* dissent, confusion, consideration, and agreement).

2.2 Engagement Estimation in HRI

Beyond the raw treatment of emotion, engagement modelling in HRI, and more generally in dialogue, has received wide research attention due to the need to keep a user engaged in joint tasks. To

date, two main aspects of user engagement have been studied in HRI [24]: First, robots should be endowed with specific features or social abilities that can increase user engagement in the conversation, such as face-tracking, performing gestures, facial expression, and voice tracking [33, 45]; Second, robots should be able to automatically recognise user engagement or disengagement during interaction. For this second class of work, Leite et al. [24] predict the intention of engagement by recognising whether users remain around a robot – which can be seen as a manifestation of spatial engagement [30]). Other studies have focused on engagement prediction by learning various features of the user *e.g.* facial expressions, gestures or postures, eye movement, and voice pitch tracking, *etc.*, [10]. A notable comprehensive study of engagement recognition with a fully autonomous robot was conducted by Ben Youssef et al. [5].

2.3 Confusion Detection

Moving from the general case of emotions and engagement to the specific cognitive state of confusion, in recent years there have been several studies aimed at defining and identifying confusion states. The theory of confusing states in emotion science is quite complex [13], with most studies on confusion detection in the area of interactive learning. D'Mello and Graesser [13] postulated that confusion is the centre of complex learning activities, such as solving difficult problems and modelling complex systems. There are numerous definitions of confusion in the literature. D'Mello and Graesser [11] summarised that confusion has been considered a bonafide emotion, an epistemic emotion, an affective state, and a mere cognitive state. When confusion is an epistemic emotion, it means that confusion is associated with impasses in the learning process when learners want to try to acquire new knowledge [28]. Cognitive disequilibrium (a mental state in which individuals encounter obstacles in their normal learning process flow [51]) can also induce stimuli confusion.

Another view is that confusion can have multiple internal states. Lodge et al. [28] defined two zones of confusion in the case of learning activities, when a learner was in cognitive disequilibrium due to an impasse in an organised learning process. In this model, the learner can at times be said to be in the zone of optimal confusion (ZOC), which is a form of productive confusion. Here, the learner is still engaged with the intention of overcoming the current impasse. However, if their confusion is persistent, the learner might instead be in a zone of sub-optimal confusion (ZOSOC); this is where confusion is unproductive, which can lead to possible frustration or boredom, with subsequent disengagement. Several papers have also focused on defining confusion, but have tended to focus on a formal model of confusion [11] and how it could relate to the learning process. For example, Arguel and Lane [1] designed two thresholds (T_a and T_b) to determine the level of confusion in a learning process. If the level of confusion is greater than T_b , indicating that the confusion is persistent, learners may be frustrated or bored. In contrast, they can engage in their current learning process when the level of confusion is below T_a . Thus, between the two thresholds, the learner is in the confusion state.

D'Mello and Graesser [11] proposed three bidirectional transitions of confusion states: the confusion-engagement transition, in

which an obstacle has been detected and the mental state has transitioned from engagement to confusion, while after the obstacle has been addressed, the state can transition back to engagement; the confusion-frustration transition, in which the obstacle cannot be continually resolved, and the mental state has transitioned from confusion to frustration. Meanwhile, if additional impasses are produced, the state is changed back to confusing; Finally, there is the frustration-boredom transition, in which the failure exists in a certain time, the learner may be disengaged, leading to boredom. However, if they have to persist in their learning task, their state will change from boredom to frustration.

Moreover, it is necessary for us in this study to understand and design how confusion can be elicited. We summarise four patterns of confusion and non-confusion induction as strategies for the stimulation of confusion [22, 46]: (a) Complex information and simple information. Lehman et al. [22] explained that complex learning is an experience full of emotions that occurs when learners are exposed to complex material, difficult issues, or indecisive decisions, such that their confusion may be triggered between positive and negative emotions [2]; (b) Contradictory information and consistent information, here people may enter into a state of uncertainty and confusion when they are exposed to contradictory information [23]; (c) Insufficient and sufficient information, here people do not receive enough information to respond to an interlocutor, and as a result, they may become confused [46]; (d) Feedback, Lehman et al. [22] designed a feedback matrix of feedback states to investigate feedback types and confusion. This matrix essentially distinguishes between correct feedback which comprises correct-positive conditions and incorrect-negative conditions, and false feedback including correct-negative and incorrect-positive conditions. From their experiment, it was seen that presentation of correct-negative feedback, *i.e.*, when the learners responded correctly but received inaccurate or negative feedback, was an effective manipulation to stimulate confusion.

Despite the above, we conclude that there has been little in terms of the creation of operationalisable models of confusion detection and modelling, and particularly in the HRI context.

3 STUDY DESIGN

Our long-term goal is to model user confusion states and apply mitigation strategies to the HRI dialogue process to alleviate confusion before user boredom or disengagement manifest. As a step along this path, we present an HRI study, which we have designed to induce confusion states in users, build a dataset of these states, and attempt to analyse these to determine whether indirect detection of confusion states might be possible.

3.1 Study Overview

For this work, we are building on an earlier pilot study in which “confusion” has been defined in situated task-oriented HRI and then invoked and studied confusion states in a remote engagement context [25]. In that earlier work, we were limited by the challenge of having users interact remotely over uncontrolled hardware (*e.g.*, microphone and camera challenges on user laptops) and the more general challenge of managing interactions remotely. Nevertheless,

in that work, we did identify certain indicators of participant confusion, and in the current study, we wish to broaden that study to provide a complete interaction scenario with a dataset that can, subject to privacy concerns, be made available for general study in language-based HRI.

In this study, we made use of a humanoid robot called Pepper. Of its many features, those that are relevant here are its onboard two high-resolution cameras as well as a 3D camera that enable the Pepper to identify movements and recognise the emotions on the faces of its interlocutors, also ability to articulate arms and head for gesticulation, and on-chest touch screen. The Pepper robot has speech recognition and dialogue available in 21 languages. For this study, the Pepper robot was configured for English. The Pepper back-end is a fully open and programmable platform built on the Naoqi framework² with comprehensive animated speech, motion, and vision modules, which are used to support our WoZ experiment.

Our study made use of a semi-spontaneous one-by-one physical face-to-face conversation between the Pepper robot and a participant in English only. The Pepper robot was controlled by a wizard. All participants were required to be able to walk into our physical laboratory. Two rooms were setup (see Figure 1): the experiment room was setup for the participants with the Pepper and some additional recording equipment. Participants were asked to remain standing in Zone 1 which is 80 cm in front of the robot, to ensure that they were close enough to Pepper yet safe for practical interaction. A high-definition (HD) webcam (Webcam 1) was placed behind the Pepper robot and aimed toward participants’ faces for collecting their facial expression. A second HD webcam (Webcam 2) was placed on the right side of the Pepper to record the body gestures of the participants. The picture on the left in Figure 1 shows the actual scene of the laboratory setting.

A researcher used the wizard room to monitor the real-time interaction of the participant and the Pepper robot in the experiment, as well as to control the Pepper robot using the WoZ4U platform [42]. The WoZ4U platform is an open-source WoZ interface that provides a graphical user interface (GUI) for the wizard to control Pepper movements, speech utterances, animated speech, gestures, *etc.* We integrated conversation scripts and developed more specific behaviours for the Pepper on the WoZ4U platform.

Consistent instructions and consent forms were provided and signed before participants attended the experiment. Live participation was designed around two interaction sessions lasting more than 15 minutes. The first session was a casual talk because most of the participants had no experience in interaction with the Pepper robot prior to participating in this experiment. To help participants adapt to the mode of human-robot dialogue, we prepared 11 interactive topics that the participant could engage in (*e.g.*, “*What is your name?*”, “*Raising your arms*”, *etc.*) as a reference so that they could feel more comfortable and confident in entering the second session. The second session was a 5 minutes task-oriented conversation between the participant and the Pepper (detailed later). The behaviours and speech of the participant were recorded in this session. After the participant finished the three tasks, they rated their confusion in a post-study survey, which was then followed by a 3-minute interview discussing this interaction (see Figure 2).

²<http://doc.aldebaran.com/2-5/naoqi/index.html>

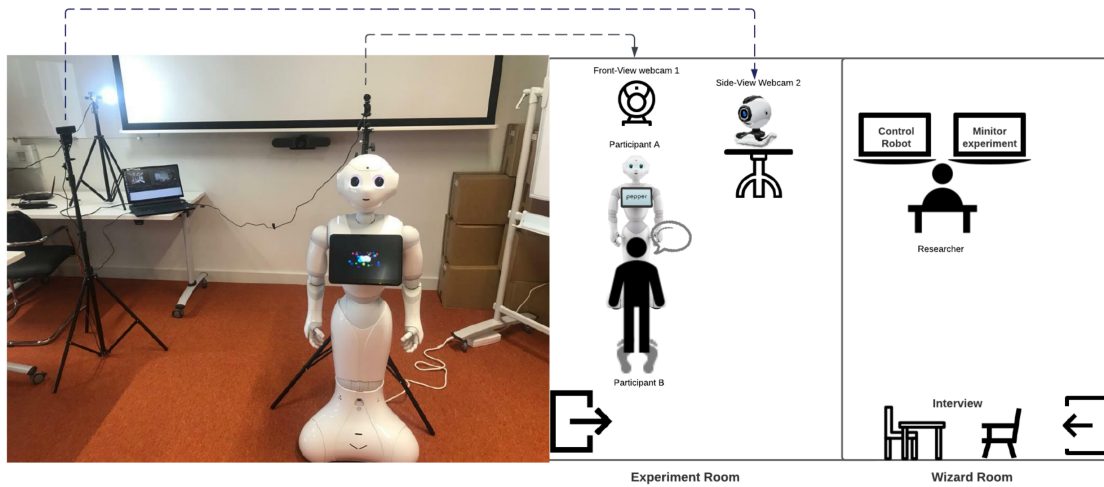


Figure 1: WoZ HRI experiment laboratory
(left: the real experiment room; right: a mock experiment room and a mock wizard room)

30 individuals participated in this study. Among them, one participant helped to first dry run this study, so this participant’s data is not analysed for research purposes. All participants were over 18 years of age, *i.e.* 5 people in the 18 - 24 age group; 23 people were in the 25 - 44 age group; and 2 people were in the 45 - 59 age group. Furthermore, they were from at least six countries and were in university programmes or industries such that they were able to have a social conversation in English. Data from 29 participants (12 female, 16 male, and one was not stated) were made available for data analysis.

3.2 Task-oriented Dialogue Design

To stimulate confusion states, two conditions were defined with different confusion stimuli: the stimuli were designed to trigger confusion in participants in Condition A; while the stimuli were designed for participants to perform a straightforward task without confusion in Condition B. Each participant was required to complete three tasks, and each task with one condition. Task 1 was a logic problem, task 2 was a word problem, and task 3 was a maths question. In order to balance the number of confusion conditions and to avoid participants having strong or persistent confusion from the sequence of tasks, the sequence of conditions for each participant could be Condition A for task 1, Condition B for task 2, and Condition A again for task 3 (*i.e.*, Condition ABA); following this participant, the sequence of conditions with the three tasks for the next participant would be task 1 with Condition B, task 2 with Condition A, and task 3 with Condition B (*i.e.*, Condition BAB).

We designed the four different confusion inductions mentioned in the literature review for each task (see Table 1). In practise, this was a small-scale study and was our first attempt to implement the confusion study on a physical robot, so we designed a short conversation for each participant, which was around 5 minutes. For example, the word problem for contradictory information in Condition A was: “there are 66 people in the playground including 28 girls, boys and teachers. How many teachers were there in total?”,

whereas the word problem for consistent information in Condition B was: “there are 5 groups of 4 students, how many students are there in the class?” (For all dialogue examples with the three tasks. Moreover, during the semi-spontaneous one-to-one conversation, we also chose one more strategy in some scenarios for confusion stimuli, for example, in Task 3 (maths questions) with complex information, some participants were not confused to show their answer correctly, which is out of our expectation, so we used the “false feedback” confusion induction to elicit confusion.

Table 1: A matrix of tasks and four pattern of confusion strategies are divided by conditions

Condition A	Condition B	Tasks *
Complex information	Simple information	Task-1,2,3
Contradictory information	Consistent information	Task-1,2,3
Insufficient information	Sufficient information	Task-1,2,3
False feedback	Correct feedback	Task-1,2,3

* Task-1,2,3: logic problem, word problem, math question

To build a more natural interaction with participants, we considered it vital that the robot possesses valid non-verbal behaviours [39]. Therefore, we designed a mapping of physical behaviours on the robot’s head, eye colours, and body gestures to align with positive and negative responses (see Figure 3).

3.3 Data Collection

The facial video data of 29 participants was labelled with a sequence of conditions such as “ABA” or “BAB”. We cropped the greeting and the end of each video. All frame data was extracted from this cropped video with conditions’ labelling noted. The image data that we extracted had 5,715 frames (3,441 frames for Condition A, 2,274 frames for Condition B). Although the image data was from facial videos, it was necessary to recognise and align faces

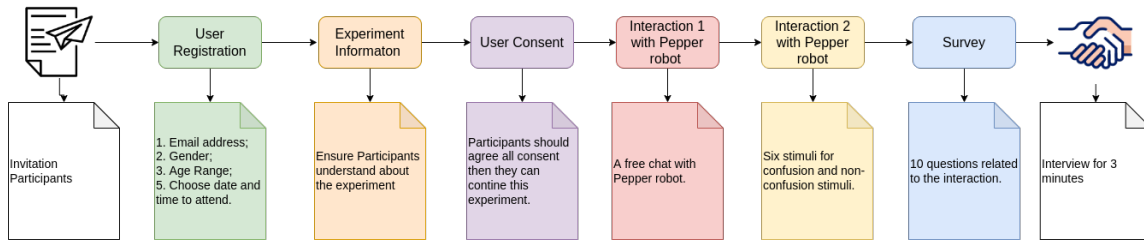


Figure 2: HRI experiment process

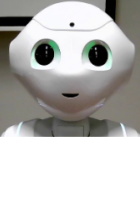
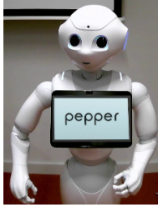
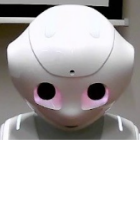

Conversational Phenomena	Communicative Behaviour		
	Face	Body	Details
Positive response			Eyes colour: green Head: face towards the participant. Body language: arms and hands are jogging swings, and head is swaying happily following the arms.
Negative response			Eyes colour: red Head: head down Body language: hands are close together in front of Pepper's body.

Figure 3: The mapping of the reaction status and visible traits for the Pepper robot

in a preprocessing data step. We approached each frame from the centre crop in a region of 224×224 pixels and we detected the face and removed the frame margins using the Multi-task Cascaded Convolutional Network (MTCNN)-based face detection algorithm [43]. Figure 4 shows a comparison of the original facial frame (left) and the centre cropped result of the same facial frame. Therefore, the result for face detection and face centre cropping was 2,945 labelled facial frames for Condition A and 1,941 facial frames for Condition B. As facial video data also included high-quality audio, we applied the FFmpeg framework to extract audio tracks for analysis. The audio data had 85 audio files (45 waveform audio (wav) files for Condition A and 40 wav files for Condition B).

Each participant completed a post-study survey with 10 questions on the Likert scale (1-5) after interacting with the robot. We designed three questions for three tasks, plus each participant evaluated their level of confusion for each confusion task. Therefore, the 29 post-study surveys were divided independently by the conditions, as we have prearranged sequences of conditions for each participant (“ABA” or “BAB”). We then combined the two independent files (one for Condition A, another for Condition B) into one file with the new “Condition” feature to mark the specific condition for subjective analysis. As there were two scores under the same

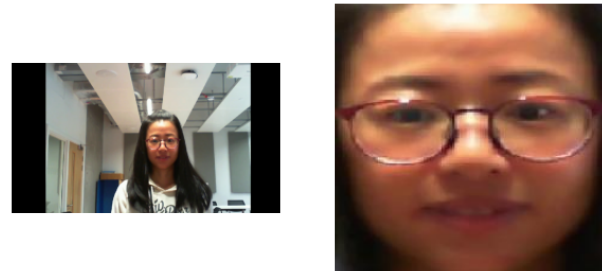


Figure 4: A comparison of facial frame (left) and aligned facial frame (right)

conditions for each questionnaire, we calculated the average of the two scores as a new parameter.

4 DATA ANALYSIS

To study the different characteristics of human behaviour under different conditions, we applied several feature extraction algorithms to our data to extract lifted features from the data. We then used these as a basis for evaluation against both post-study self-rating results and our experimental conditions.

4.1 Visual Data Measurement

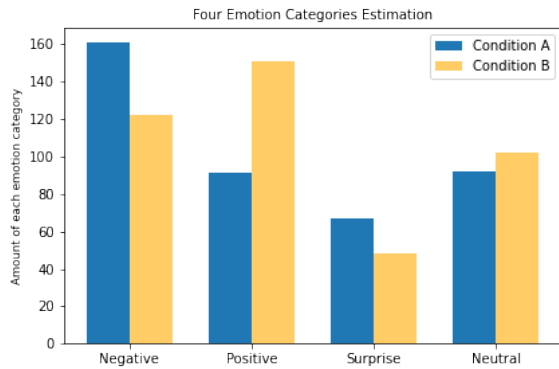
We applied a facial emotion detection algorithm to our preprocessed frame data; this algorithm used the MobileNet architecture and was trained on the AffectNet dataset [18, 31, 43]. This resulted in estimates of each of the seven emotion categories (neutral, happy, sad, surprise, fear, anger, and disgust) for each frame. For the facial emotion analysis, a Chi-Square test for independence (with Yates’ Continuity Correction) indicated significant association between conditions and seven emotion categories, $\chi^2(1, n = 4886) = 25.01, p < 0.05, phi = 0.07$.

Furthermore, table 2 shows the number of each of the seven categories of emotions grouped by conditions and normalised by total detection. We noticed that the number of fear emotions is much higher than the other six emotions. On investigation, we see this as a limit or bias in the algorithm and subsequently removed the count of fear labels from further analysis. Furthermore, we summed the quantity of negative emotions (anger, disgust, and sadness) grouped by the two conditions, and note that the number of negative emotions in Condition A is notably greater than in Condition B. Consequently, the number of predicted positive emotions

Table 2: The result of emotion estimation grouped by Condition A and Condition B

Condition	Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral	Overall
A	40	62	1511	59	91	67	92	1922
B	19	46	1503	57	151	48	102	1926

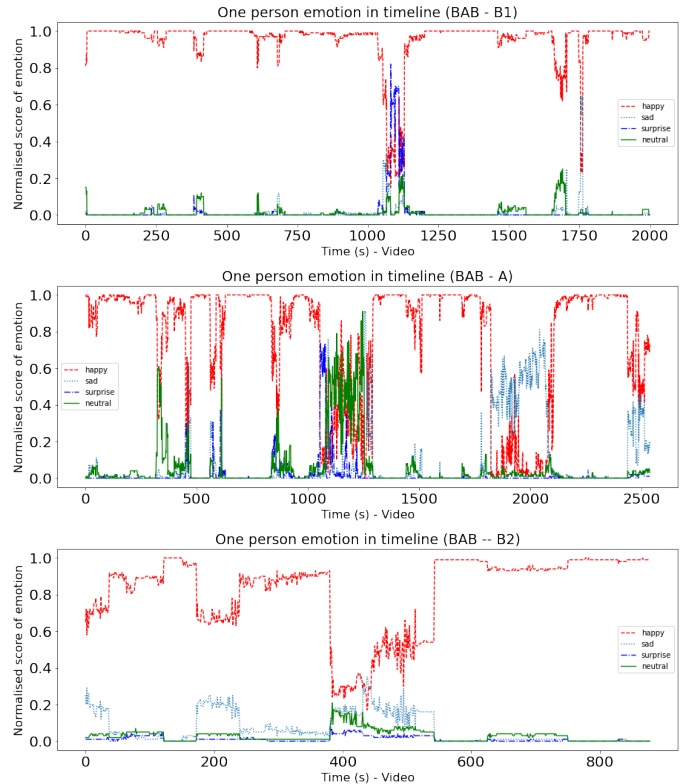
(happiness) results for Condition A is less than that for Condition B. Similarly, surprise, that is an emotion either a negative or positive emotion in different contexts [49], was higher in Condition A than in Condition B. Finally, the predicted results of the neutral results for Condition A are less than those for Condition B (see Figure 5 for a summary of these aggregated results).

**Figure 5: The four emotion categories grouped by conditions**

As a form of error analysis, we applied the Facial Expression Recognition (FER) open-source framework to the preprocessed videos. The FER framework is built with the MTCNN for facial recognition [52] and an emotion classifier [3] that has been trained on the FER-2013 emotion dataset [15]. To illustrate, we plotted the four primary changes in emotions following a sequence of conditions for the three tasks (*i.e.* BAB). Figure 6 shows that the emotion “happy” of the participant was dominated most of the time in the two instances of Condition B, with scores of “happy” approaching 1. Whereas the proportion of “happy” decreased, while each of “sad”, “surprise”, and “neutral” became dominant in periods of Condition A.

Turning to eye gaze, we applied a state-of-the-art eye gaze estimator, trained on the ETH-XGaze dataset [53], to predict pitch and yaw angles for each preprocessed facial frame. We summed the absolute two angles as a new feature for statistical analysis, since a human has different angles of direction corresponding to positive or negative values of pitch and yaw, leading to the sum of values being 0. An independent-samples *t*-test was conducted to compare the normalised angles of pitch and yaw for eye gaze under the two experimental conditions. A significant difference was found in the normalised pitch and yaw for eye gaze ($M = 0.38, SD = 0.14$ for Condition A, $M = 0.40, SD = 0.14$ and Condition B), $t(2587) = -1.99, p < 0.05, d = -0.08$.

To investigate the overall trend, we also investigated individual traces as a form of error analysis. Figure 7 shows the fluctuations of the pitch and yaw angle for the two time periods labelled Condition

**Figure 6: The emotional changes for one participant during the three tasks with conditions (BAB)**

B and the one time period of Condition A for one participant. For the fluctuations following a sequence of conditions (BAB), we can see that the average area of two angles in Condition A is greater than the average area of two angles in Conditions B instances.

For head-pose estimation, the model that we applied is designed using CNNs, dropout, and adaptive gradient methods [37], and trained on three popular datasets (*i.e.*, the Prima head-pose dataset, the Annotated Facial Landmarks in the Wild (AFLW) and the Annotated face in the Wild (AFW) dataset) [16, 21, 54]. The predicted results are the three angles of pitch, yaw and roll. Again we calculated an aggregate value of the normalised three absolute angles as a new variable. The result of an independent-samples *t*-test, however, showed that there was, no significant difference in the angles of roll, pitch and yaw for head pose ($M = 0.26, SD = 0.13$ for Condition A, $M = 0.26, SD = 0.13$ for Condition B), $t(5713) = -0.49, p = 0.62, d = -0.01$. Nevertheless, given the independence of these variables in specific social behaviours, we also analysed the three independent results (*i.e.* pitch, yaw, and roll angles) with two conditions. The result for the normalised pitch angles with two

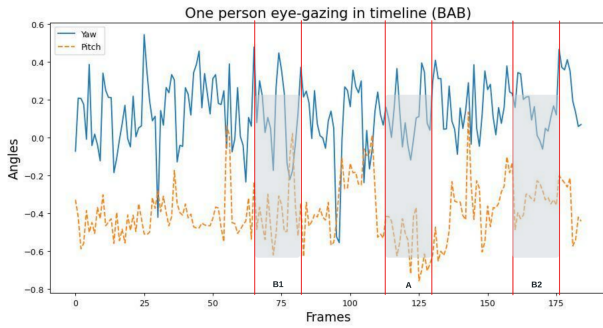


Figure 7: Changes of one person’s pitch and yaw angles of eye-gazing in timeline with conditions (BAB)

conditions is that there is a significant difference in the angles of pitch ($M = 0.37, SD = 0.17$ for Condition A, $M = 0.41, SD = 0.17$ for Condition B), $t(5713) = -7.16, p < 0.05, d = -0.19$; Similarly, the result for the normalised yaw angles with two conditions is that a significant difference in the angles of yaw ($M = 0.54, SD = 0.11$ for Condition A, $M = 0.63, SD = 0.12$ for Condition B), $t(5713) = -31.14, p < 0.05, d = -0.82$; However, the result for the normalised roll angles with two conditions is that there was no significant difference in the angles of roll for head pose ($M = 0.56, SD = 0.14$ for Condition A, $M = 0.55, SD = 0.12$ for Condition B), $t(5713) = 0.36, p = 0.72, d = 9.49e - 03$.

4.2 Audio Data Measurement

The phenomenon of silence during conversations has been analysed in numerous pragmatic studies [35]. There are two types of silence with respect to the specific state of the interlocutor: intentional silence, where the interlocutor refuses to respond to a speaker; and unintentional silence where the interlocutor psychologically cannot respond to a speaker [20, 35]; both, however, can be relevant in the case of confusion and disengagement. Therefore, we calculate the duration of silence for each audio sample for each of our conditions. An independent-samples t-test was conducted to compare the normalised silence duration time between the two conditions. Here, a significant difference was found between the normalised silence duration time and the two conditions ($M = 0.45, SD = 0.23$ for Condition A, $M = 0.27, SD = 0.19$ for Condition B), $t(83) = 3.94, p < 0.05, d = 0.86$.

For illustration purposes, we plotted the normalised silence duration time of the two conditions (see Figure 8), showing that the silence duration values of Condition A form a more discrete distribution than those of Condition B. Meanwhile, for an individual observation, Figure 9 shows the normalised duration of silence for one participant in the three tasks performed with the BAB sequence conditions; here the duration time of silence for Condition A is obviously longer than for both Conditions B.

4.3 Subjective Measurement

We analysed the post-study survey scores with the two independent groups split by two controlled conditions for the stimuli. Two

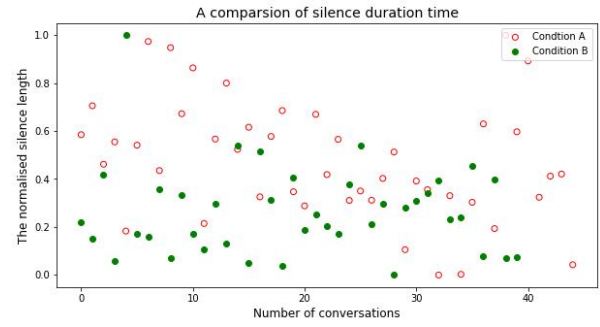


Figure 8: Plotted silence duration time grouped by conditions

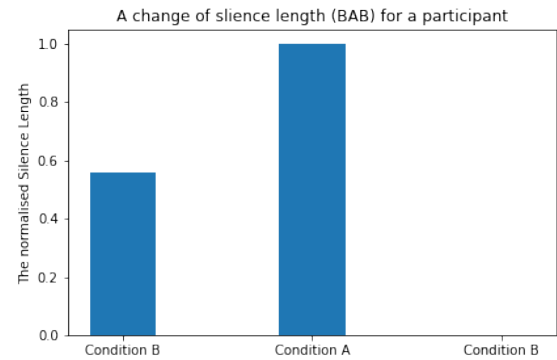


Figure 9: A change of silence duration with conditions (BAB)

statistical questions were investigated: (1) the relationship between the three task-centric confusion sub-question scores and the two conditions; (2) whether there was a significant relationship between the average self-reported confusion scores for the three tasks and the two conditions.

An Mann-Whitney U-test was conducted for first question. The results of the first question with the three sub-questions: (1) the confusion levels for task 1 (a logic problem) with Condition A ($Mdn = 0, IQR = 2.00$) were not significantly higher than those with Condition B ($Mdn = 1, IQR = 3.00$), $U = 404.50, Z = -0.27, p = 0.80, r = 0.04$. (2) The confusion levels for task 2 (a word problem) with Condition A ($Mdn = 1.00, IQR = 2.00$) was not significantly higher than those with Condition B ($Mdn = 0, IQR = 2.00$), $U = 417.50, Z = -0.05, p = 0.97, r = 0.07 - e1$. And (3) the confusion levels for task 3 (a maths question) with Condition A ($Mdn = 0, IQR = 3.00$) was not significantly higher than those with Condition B ($Mdn = 1, IQR = 2.00$), $U = 442, Z = 0.36, p = 0.73, r = 0.05$.

An independent-sample t-test was performed for the second question, which showed that there was no significant difference in normalised average confusion levels for the three tasks performed and the two conditions ($M = 0.37, SD = 0.29$ for Condition A, $M = 0.38, SD = 0.35$ for Condition B), $t(56) = -0.10, p = 0.92, d = -0.03$.

5 DISCUSSION

Given the results just presented, we can make several observations. (1) Participants were not necessarily aware of being confused when

presented with confusion stimuli. (2) Participant's emotions were more negative and more surprised in confusion conditions than non-confusion. (3) For the changes of emotions in the participant on the timeline of the HRI experiment, the positive or neutral emotion might be a main emotion category in the task-oriented HRI and the negative emotion might increase in confusion conditions. (4) Participants' ranges of eye gaze angles were less in confusion than in non-confusion situations. (5) For the changes of eye-gaze angles in the timeline with different conditions, their gaze range fluctuated more following the interaction with Condition A than after interacting with Condition B. (6) It was found that there is a strong correlation of participants' pitch angles of head pose and yaw angles of head pose respectively, in addition to their roll angles of head pose, with confusion or non-confusion, although there was no strong correlation of their ranges of absolute summed three angles of head pose with confusion or non-confusion. (7) The silence duration time was longer in confusion than in non-confusion.

Compared to the existing human-like avatar interaction work, in our HRI study, although the Pepper lacks anthropomorphic facial expression, the robot has advanced body language and appropriate automatic animated speech. Meanwhile, the quality of data collection is guaranteed as we controlled most of the variables in the experiment environment. In addition, in a 3-minute interview, most of the participants were surprised that the robot has high-tech social interaction skills and friendly behaviour.

5.1 Limitations

The study was built on an earlier pilot study that was conducted with a remote avatar-based setting; however, this study had a number of limitations worth mentioning. (1) 25 out of the 29 participants had a technical background in computer science; therefore, we expect biases in population interactions relative to the general population. This is a controlled lab study, although we recruited people in public, it seems those kinds of people would like to register the experiment [14, 17]. (2) There was no control over the conversation boundaries to reflect different states of confusion (*i.e.* productive confusion or unproductive confusion) in these short conversations with confusion stimuli. (3) As it is an early study within a complete study programme, we have focused on designing the situated task-oriented HRI to track interlocutor confusion, and as a result it lacks a third party (*e.g.* annotators or experts) to further test our design for high-fidelity confusion innovation in a plausible setting. (4) In our controlled HRI experiment, the participants were assigned each task with only one type and one condition, and different types of tasks that can raise the participants' confusion are not included in this study. (5) We note that post-study surveys have a risk and noisy that we cannot ensure that participants can remember their confusion, it is possible to impact results of subjective measurement. However, in this study, the influence of the participant's memory is limited, since each task approximately lasts only 1.5 minutes. Therefore, we are confident that most participants should in principle remember their confusion for each task.

5.2 Future Study

In terms of further study, the future experiment design, based on our experience and data analysis, will focus on the word problem task

as these are more straightforward to design with four inductions for confusion stimuli. We plan to use this task with four confusion inductions for each participant to reduce variables from mixed confusion inductions in one task. Although we do intend to lengthen the period and complexity of individual tasks, users will have time to directly after each task rate their confusion score and a brief rest to prepare them for the next task.

As for the data collection and analysis, the sample size of this pilot study was only 29 samples, in the next study, we will increase the sample size (target: 60) to generate more stable and reliable results. However, since we will be building the experiment in broadly the same way, we do hope to be able to compare results across studies as a form of validation. Furthermore, recognition of body posture and speech emotion are two primary avenues for further feature analysis in multiple confusion states. To further verify the dialogue strategies for confusion stimuli can elicit confusion successfully, and also to ensure the more reliable results of data analysis, an annotation schema of confusion and non-confusion will be designed for annotators to annotate different confusion conditions on our future data. More importantly, though, we recognise that being able to generalise from specific user studies such as this is very challenging, and as such is a major component of future work may be to establish generalised abstracted models that can be applied across different social and technological settings. Within our group, we are also working on developing dynamic planning policies which can be adjusted in the case of confusion detection.

6 CONCLUSION

In conclusion, this paper illustrated with a controlled study that even when users are not aware of being in a confused state, they present different interaction behaviours which may in principle be detected by automated systems such as social robots. We believe that these results do validate that behavioural differences between visual and speech behaviour are present in confusing and non-confusing situations. Meanwhile, this study not only motivates our further research in this domain, but also has the potential to increase the social task-oriented capabilities of dialogue-equipped robots in the medium to long term. This is our first study on modelling confusion states in a situated HRI task-oriented dialogue setting. Nevertheless, we see it as a firm foundation for further situated dialogue investigation for HRI, and in particular, where we focus on enhancing engagement through preemptive anticipation of disengagements. In future work, we aim to expand the feature analysis, generalise the abstracted models to be applicable across platforms, and develop generalised dialogue strategies/policies to provide additional and clarification information to users to assist them in joint task performance.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

- [1] Amaël Arguel and Rod Lane. 2015. Fostering deep understanding in geography by inducing and managing confusion: An online learning approach. *ASCILITE 2015 - Australasian Society for Computers in Learning and Tertiary Education, Conference Proceedings* (2015), 374–378.
- [2] Amaël Arguel, Lori Lockyer, Ottmar V. Lipp, Jason M. Lodge, and Gregor Kennedy. 2017. Inside Out: Detecting Learners' Confusion to Improve Interactive Digital Learning Environments. *Journal of Educational Computing Research* 55, 4 (2017), 526–551. <https://doi.org/10.1177/0735633116674732>
- [3] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time Convolutional Neural Networks for Emotion and Gender Classification. <https://doi.org/10.48550/ARXIV.1710.07557>
- [4] Christoph Bartneck, Tony Belpaeme, Friederike Eyssele, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *References*. Cambridge University Press. 209–246 pages.
- [5] Atef Ben Youssef, Chloé Clavel, Slim ESSID, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. 464–472. <https://doi.org/10.1145/3136755.3136814>
- [6] Mohamed Sahbi Benlamine and Claude Frasson. 2021. Confusion Detection Within a 3D Adventure Game. In *Intelligent Tutoring Systems*, Alexandra I. Cristea and Christos Troussas (Eds.). Springer International Publishing, Cham, 387–397.
- [7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (2008), 335–359.
- [8] O Celiktutan, S Skordos, and Hatice Gunes. 2017. Multimodal Human-Human-Robot Interactions (MHRI) Dataset for Studying Personality and Engagement. (2017). <https://doi.org/10.17863/CAM.13433>
- [9] Jeffrey F. Cohn. 2007. Foundations of Human Computing: Facial Expression and Emotion. In *Artificial Intelligence for Human Computing*, Thomas S. Huang, Anton Nijholt, Maja Pantic, and Alex Pentland (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–16.
- [10] M. A. Dewan, M. Murshed, and F. Lin. 2018. Engagement detection in online learning: a review. *Smart Learning Environments* 6 (2018), 1–20.
- [11] Sidney D'Mello and Art Graesser. 2014. Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica* 151 (2014), 106–116. <https://doi.org/10.1016/j.actpsy.2014.06.005>
- [12] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2014), 153–170. <https://doi.org/10.1016/j.learninstruc.2012.05.003>
- [13] Sidney K. D'Mello and Arthur C. Graesser. 2000. Chapter 15 - Confusion. In *Intellectual Property Law Q&A (2nd ed.)*, Routledge-Cavendish (Ed.). 9781843141495, 289–310. <https://doi.org/10.4324/9781843141495>
- [14] Kerstin Fischer. 2021. Effect Confirmed, Patient Dead: A Commentary on Hoffman; Zhao's Primer for Conducting Experiments in HRI. *J. Hum.-Robot Interact.* 10, 1, Article 9 (feb 2021), 4 pages. <https://doi.org/10.1145/3439714>
- [15] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. Challenges in Representation Learning: A report on three machine learning contests. <https://doi.org/10.48550/ARXIV.1307.0414>
- [16] Nicolas Gourier, D. Hall, and J. Crowley. 2004. Estimating Face orientation from Robust Detection of Salient Facial Structures.
- [17] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 2-3 (2010), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- [18] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR abs/1704.04861* (2017). arXiv:1704.04861 <http://arxiv.org/abs/1704.04861>
- [19] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural Responses to Robot Conversational Failures. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 53–62.
- [20] Dennis Kurzon. 1998. *Discourse of Silence*. John Benjamins. <https://www.jbe-platform.com/content/books/9789027282606>
- [21] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2011. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2144–2151. <https://doi.org/10.1109/ICCVW.2011.6130513>
- [22] Blair Lehman, Sidney D'Mello, and Art Graesser. 2012. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education* 15, 3 (2012), 184–194. <https://doi.org/10.1016/j.iheduc.2012.01.002>
- [23] Blair A. Lehman, Sidney K. D'Mello, and Arthur C. Graesser. 2013. Who Benefits from Confusion Induction during Learning? An Individual Differences Cluster Analysis. In *AIED*.
- [24] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scarsellati. 2015. Comparing Models of Disengagement in Individual and Group Interactions. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) (HRI '15). Association for Computing Machinery, New York, NY, USA, 99–105. <https://doi.org/10.1145/2696454.2696466>
- [25] Na Li, John D Kelleher, and Robert Ross. 2021. Detecting Interlocutor Confusion in Situated Human-Avatar Dialogue: A Pilot Study. In *25th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021)* University of Potsdam, Germany. <https://doi.org/10.21427/bsd0-7326>
- [26] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. DECN: Dialogical emotion correction network for conversational emotion recognition. *Neurocomputing* 454 (2021), 483–495. <https://doi.org/10.1016/j.neucom.2021.05.017>
- [27] Jing Liang, Yu-Qing Zou, Si-Yi Liang, Yu-Wei Wu, and Wen-Jing Yan. 2021. Emotional Gaze: The Effects of Gaze Direction on the Perception of Facial Emotions. *Frontiers in Psychology* 12 (2021). <https://doi.org/10.3389/fpsyg.2021.684357>
- [28] Jason M. Lodge, Gregor Kennedy, Lori Lockyer, Amaël Arguel, and Mariya Pachman. 2018. Understanding Difficulties and Resulting Confusion in Learning: An Integrative Review. *Frontiers in Education* 3 (2018). <https://doi.org/10.3389/feeduc.2018.00049>
- [29] Isabelle M. Menne and Birgit Lugin. 2017. In the Face of Emotion: A Behavioral Study on Emotions Towards a Robot Using the Facial Action Coding System. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) (HRI '17). Association for Computing Machinery, New York, NY, USA, 205–206. <https://doi.org/10.1145/3029798.3038375>
- [30] M.P. Michalowski, S. Sabanovic, and R. Simmons. 2006. A spatial model of engagement for a social robot. In *9th IEEE International Workshop on Advanced Motion Control, 2006*, 762–767. <https://doi.org/10.1109/AMC.2006.1631755>
- [31] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *CoRR abs/1708.03985* (2017). <https://doi.org/10.1109/TAFFC.2017.2740923>
- [32] Cecilia G. Morales, Elizabeth J. Carter, Xiang Zhi Tan, and Aaron Steinfeld. 2019. Interaction Needs and Opportunities for Failing Robots. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) (DIS '19). Association for Computing Machinery, New York, NY, USA, 659–670. <https://doi.org/10.1145/3322276.3322345>
- [33] Lilia Moshkina, Susan Trickett, and J. Gregory Trafton. 2014. Social Engagement in Public Places: A Tale of One Robot. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) (HRI '14). Association for Computing Machinery, New York, NY, USA, 382–389. <https://doi.org/10.1145/2559636.2559678>
- [34] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2009. Head Pose Estimation in Computer Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 4 (2009), 607–626. <https://doi.org/10.1109/TPAMI.2008.106>
- [35] Naoki Ohshima, Keita Kimijima, Junji Yamato, and Naoki Mukawa. 2015. A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 325–330. <https://doi.org/10.1109/ROMAN.2015.7333677>
- [36] Mariya Pachman, Amaël Arguel, Lori Lockyer, Gregor Kennedy, and Jason M Lodge. 2016. *Eye tracking and early detection of confusion in digital learning environments: Proof of concept*. Technical Report 6, 32 pages.
- [37] Massimiliano Patacchiola and Angelo Cangelosi. 2017. Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods. *Pattern Recognition* 71 (2017), 132–143. <https://doi.org/10.1016/j.patcog.2017.06.009>
- [38] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125. <https://doi.org/10.1016/j.inffus.2017.02.003>
- [39] Vignesh Prasad, Ruth Stock-Homburg, and Jan Peters. 2020. Advances in Human-Robot Handshaking. In *Social Robotics*, Alan R. Wagner, David Feil-Seifer, Kerstin S. Haring, Silvia Rossi, Thomas Williams, Hongsheng He, and Shuzhi Sam Ge (Eds.). Springer International Publishing, Cham, 478–489.
- [40] Chowdhury Mohammad Masum Refat and Norsinnira Zainul Azlan. 2019. Deep Learning Methods for Facial Expression Recognition. In *2019 7th International Conference on Mechatronics Engineering (ICOM)*, 1–6. <https://doi.org/10.1109/ICOM47790.2019.8952056>
- [41] L. Riek. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. In *HRI 2012*.
- [42] Finn Rietz, Alexander Sutherland, Suna Bensch, Stefan Wermter, and Thomas Hellström. 2021. WoZ4U: An Open-Source Wizard-of-Oz Interface for Easy, Efficient and Robust HRI Experiments. *Frontiers in Robotics and AI* 8 (2021). <https://doi.org/10.3389/frobt.2021.668057>

- [43] Andrey V. Savchenko. 2021. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *CoRR* abs/2103.17107 (2021). arXiv:2103.17107 <https://arxiv.org/abs/2103.17107>
- [44] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to Look: A Study of Human-Robot Engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) (IUI '04). Association for Computing Machinery, New York, NY, USA, 78–84. <https://doi.org/10.1145/964442.964458>
- [45] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1 (2005), 140–164. <https://doi.org/10.1016/j.artint.2005.03.005>
- [46] P. Silvia. 2010. Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts* 4 (2010), 75–80.
- [47] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. *Frontiers in Robotics and AI* 7 (2020), 145. <https://doi.org/10.3389/frobt.2020.532279>
- [48] Nhan Tran, Kai Mizuno, Trevor Grant, Thao Phung, Leanne Hirshfield, and Tom Williams. 2020. Exploring Mixed Reality Robot Communication Under Different types of Mental Workload. *International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interaction* 3 (2020). <https://doi.org/10.31219/osf.io/f3a8c>
- [49] Pascal Vrticka, Lara Lordier, Benoit Bediou, and David Sander. 2014. Human amygdala response to dynamic facial expressions of positive and negative surprise. *Emotion* 14 1 (2014), 161–9.
- [50] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A. Neerincx. 2014. Robot Mood is Contagious: Effects of Robot Body Language in the Imitation Game (AAMAS '14). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 973–980.
- [51] Diyi Yang, Robert E Kraut, Carolyn P Rosé, and Ros ´ Rosé. 2015. *Exploring the Effect of Student Confusion in Massive Open Online Courses*. Technical Report. <http://www.katyjordan.com/MOOCproject.html>
- [52] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (oct 2016), 1499–1503. <https://doi.org/10.1109/lsp.2016.2603342>
- [53] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. *CoRR* abs/2007.15837 (2020). https://doi.org/10.1007/978-3-030-58558-7_22
- [54] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2879–2886. <https://doi.org/10.1109/CVPR.2012.6248014>

Table 3: Examples of conversation scripts for confusion stimuli

Tasks*	Confusion Pattern*	Condition A	Condition B
1	(1) (4)	R*: Suppose Anna’s mother admires Anna, Anna admires her mother, everyone admires her mother, so everyone admires Anna, right? P*: <user-response> R: Does it make sense Anna’s friend admires Anna but her mother, doesn’t it? P: <user-response> R: Thank you for your answer.	R: Suppose everyone over the age of 30 is a liar, William is a liar, so the question is, is William over 30? P: <user-response> R: Do you agree that not everyone under 30 is not a liar? P: <user-response> R: Great, you are correct.
2	(3) (4)	R: There are 66 people in the playground including 28 girls, boys and teachers. How many teachers were there in total? P: <user-response> R: Please try again. P: <user-response> R: Thank you for your answer.	R: There are 5 groups of 4 students, how many students are there in the class? P: <user-response> R: You are correct. R: Each group has 2 pairs of scissors, how many pairs of scissors are there in total? P: <user-response> R: Well done, you are so smart.
3	(1) (4)	R: If $x = 4$ and $x + b + \log(1) = 10$, the question is, is $b = 6$ or $b = 12$? P: <user-response> R: Please try again. P: <user-response> R: Sorry, maybe this question is too difficult.	R: If $x = 4$ and $x + b = 10$, the question is, is b equal 12? P: <user-response> R: Great, you are correct.

* R: Robot; P: Participant * (1) Complex information and simple information; (2) Contradictory information and consistent information; (3) Insufficient and sufficient information; (4) feedback. * Task 1: logic problem, Task 2: word problem, Task 3: math question

Table 4: User Survey for HRI Study (5-point Likert Scales)

No.	Questions
1	Did you enjoy talking to Pepper overall?
2	Was the conversation with Pepper fluent?
3	Was the conversation with Pepper easy?
4	Was the conversation with Pepper frustrating?
5	Was the conversation with Pepper boring?
6	Did you feel confused most of the time talking with Pepper?
7	Did you feel confused when you answered the logical questions to Pepper (including Pepper’s responses may make you confused)?
8	Did you feel confused when you answered the word problems to Pepper (including Pepper’s responses may make you confused)?
9	Did you feel confused when you answered the Mathematics questions to Pepper (including Pepper’s responses may make you confused)?
10	Did you want to give up this conversation with Pepper?